1984-04-01

Test-Retest Reliability of a Formula-Scored Multiple-Choice Test

Weitzman, R.A.

Sage Journals

Weitzman, R. A. "Test-retest reliability of a formula-scored multiple-choice test." Psychological reports 54.2 (1984): 419-425.
http://hdl.handle.net/10945/60662

Psychological Reports, 1984, 54, 419-425.

TEST-RETEST RELIABILITY OF A FORMULA-SCORED MULTIPLE-CHOICE TEST[1]

R. A. WEITZMAN

Naval Postgraduate School

Summary.—In an ideal multiple-choice test, defined as a multiple-choice test containing only items with options that are all equally guessworthy, the probability of guessing the correct answer to an item is equal to the reciprocal of the number of the item's options. This article presents an asymptotically exact estimator of the test-retest reliability of an ideal multiple-choice test. When all test items have the same number of options, computation of the estimator requires, in addition to the number of options per item, the same information as computation of the Kuder-Richardson Formula 21: the total number of items answered correctly on a single testing occasion by each person tested. Both for ideal multiple-choice tests and for nonideal multiple-choice tests for which the *average* probability of guessing the correct answer to an item is equal to the reciprocal of the number of options per item, Monte Carlo data show that the estimator is considerably more accurate than the Kuder-Richardson Formula 21 and, in fact, is very nearly exact in populations of the order of 1000 persons.

The test-retest reliability of a test, which is defined as the correlation between identical versions of the test taken by the same people on independent occasions, has served psychologists and educators more as an ideal concept than as an actual quantity. The difficulty, of course, is that, because assuring independent testing occasions is impossible, test-retest reliabilities have not been computable in practice. Other measures of test reliability exist; but, as Guttman (1945) has shown, all measures of test reliability in common use are necessarily lower than the test-retest measure. Each commonly used measure of test reliability, moreover, has a distinct interpretation which for a test of standard length is unambiguous when the measure is near unity. What is not generally recognized, however, is that, when any of these measures of the reliability of a test is low, unambiguous interpretation is impossible unless the test-retest reliability of the test is known. Although high values of Kuder-Richardson Formulas 20 and 21 (Kuder & Richardson, 1937), for example, indicate that the items of a standard-length test are homogeneous with respect to content (Gulliksen, 1950, Chapter 16), low values of these statistics do not necessarily indicate that the items are heterogeneous. Any measure of the reliability of a test may be low simply because the test-retest reliability of the test is low; without knowledge of the test-retest reliability, this possibility cannot be ruled out. For this reason, precise knowledge of the test-retest reliability

[1]Requests for reprints should be sent to the author, Department of Administrative Sciences, Naval Postgraduate School, Monterey, California 93943.

of a test may be needed even when other measures of the test's reliability are available.

This article will develop a method for estimating the test-retest reliability of a multiple-choice test from the results of a single testing occasion. Based on the assumption that every person answers every test item, the estimators derived will be, with respect to the number of persons tested, asymptotically exact whenever all the options of every test item are equally guessworthy. Although this condition may seem natural enough to some people, there are very likely no conditions affecting guessing that everyone would agree are generally desirable or even of particularly common occurrence. Sound reasons exist for trying to achieve this condition, however, and this condition *is* testable (Weitzman, 1970). As Monte Carlo data show, moreover, even when this condition does not hold, the estimators derived are likely to be highly accurate.

A multiple-choice item with options that are all equally guessworthy has been called *ideal*, and a multiple-choice test that contains only ideal items has been called an ideal multiple-choice test (Weitzman, 1970). The following development applies to ideal multiple-choice tests scored by the standard correlation for guessing (the number right minus $1/(A-1)$ times the number wrong for A-choice items). This development brings multiple-choice testing and, particularly, the standard correction for guessing into the context of mental-test theory. In standard treatments of mental-test theory (Lord & Novick, 1967, Ch. 14), virtually the only reference to multiple-choice tests is in connection with scoring formulas. These formulas show that true-score estimates depend on the number of choices per item. The error variance, however, also depends on this number. Although intuitively obvious, this dependence has never been formally incorporated into reliability estimation prior to the development presented below.

Development of the Formulas

In the classical development (Gulliksen, 1950, Ch. 2; Lord & Novick, 1968, Ch. 3), the reliability of a test is defined as the correlation between the test and another "parallel" test having equal true scores and the same error variance. Defining reliability in reference to only a single test as the correlation between two independent replications of the test, the development here makes no assumptions of parallelism. According to this definition,

$$\rho_p(X_{pt}, X_{pt'}) = [\sigma_p(X_{pt}, X_{pt'})]/[\sigma_p(X_{pt})\sigma_n(X_{pt'})] , \qquad [1]$$

where X_{pt} is the score of person p on replication (trial) t. This is the definition that Guttman (1945) used to show that many common single-test estimators of reliability are actually lower bounds. The estimators developed here will tend, by contrast, to be exact in large samples of people. To proceed from the definition of Equation [1] to formulas involving only scores observed on a

single replication requires use of a principle that Guttman (1945) called "convergence in the mean": A sample mean (E_p) tends to be equal to its expected value (E_tE_p) as the sample size approaches infinity ($N \to \infty$). Aside from straightforward algebra, the development that follows will thus involve the substitution of E_pE_t for E_p. The results, accordingly, will be asymptotic with respect to sample size.

The measurement model used in the development is the identity that breaks a deviation score into true and error components:

$$(X_{pt} - \mu) = (T_p - \mu) + (X_{pt} - T_p) \ , \qquad [2]$$

with $E_p(T_p) = \mu$ and $E_t(X_{pt}) = T_p$. Substitution from this model for $(X_{pt} - \mu)$ in the covariance $\sigma_p(X_{pt}, X_{pt'})$ of Equation [1] produces the sum of the following four terms: $E_p(T_p - \mu)^2$, $E_p(T_p - \mu)(X_{pt'} - T_p)$, $E_p(T_p - \mu)(X_{tp} - T_p)$, and $E_p(X_{pt} - T_p)(X_{pt'} - T_p)$. Replacement of E_p by E_pE_t in the last three of these terms shows them all to be equal to zero since $E_t(X_{pt} - T_p) = E_t(X_{pt'} - T_p) = 0$ and, the replications being independent, $E_t(X_{pt} - T_p)(X_{pt'} - T_p) = 0$. Remaining only is the first term, the true score variance, σ_T^2, so that

$$\sigma_p(X_{pt}, X_{pt'}) \underset{N\to\infty}{=} \sigma_T^2 \ , \qquad [3]$$

the result well-known for parallel tests. Following the same procedure for the variance $\sigma_p^2(X_{pt})$ produces the sum of three terms: $E_p(X_{pt} - T_p)^2$, $2E_p(X_{pt} - T_p)(T_p - \mu)$, and $E_p(T_p - \mu)^2$. Replacement of E_p by E_pE_t in the second term shows it to be equal to zero since $E_t(X_{pt} - T_p) = 0$, and corresponding replacement in the first term shows it to be equal to $E_p\sigma_t^2(X_{pt})$, the mean error variance over people. Since the third term is the true-score variance σ_T^2,

$$\sigma_p^2(X_{pt}) \underset{N\to\infty}{=} \sigma_T^2 + E_p\sigma_t^2(X_{pt}) \ . \qquad [4]$$

This result does not depend on the replication t or t', and so

$$\rho_p(X_{pt}, X_{pt'}) \underset{N\to\infty}{=} \sigma_T^2/[\sigma_T^2 + E_p\sigma_t^2(X_{pt})] \qquad [5]$$

or

$$\rho_p(X_{pt}, X_{pt'}) \underset{N\to\infty}{=} [\sigma_p^2(X_{pt}) - E_p\sigma_t^2(X_{pt})]/[\sigma_p^2(X_{pt})] \ , \qquad [6]$$

which would be computable from the test scores of a single replication if the error variance $E_p\sigma_t^2(X_{pt})$ were computable.

The error variance $E_p\sigma_t^2(X_{pt})$ is in fact computable if the test is multiple-choice with A options for each item so that the error is due to guessing with

the probability of a correct guess equal to 1/A. If R_{pt} is the number of correct responses of person p on replication t, then for an n-item test

$$E_t(R_{pt}) = T_p + [(n - T_p)/A] \quad [7]$$

(the expected number correct is equal to the number known T_p plus the expected number guessed $(n - T_p)/A$) and

$$\sigma_t^2(R_{pt}) = (n - T_p)(1/A)[1 - (1/A)] \ , \quad [8]$$

the binomial variance for the unknown $(n - T_p)$ items. Substitution of X_{pt} for T_p in Equation [7] leads to the standard correction for guessing

$$X_{pt} = R_{pt} - [(n - R_{pt})/(A - 1)] \ , \quad [9]$$

so that $E_t(X_{pt}) = T_p$ and $\sigma_t^2(X_{pt}) = [A/(A - 1)]^2 \sigma_t^2(R_{pt})$ or, from Equation [8],

$$\sigma_t^2(X_{pt}) = (n - T_p)/(A - 1) \ , \quad [10]$$

a result presented by Weitzman (1968) and Lord and Novick (1968, p. 308). Since $E_t(X_{pt}) = T_p$, then,

$$E_p\sigma_t^2(X_{pt}) = E_pE_t[(n - X_{pt})/(A - 1)] \ , \quad [11]$$

or on replacement of E_pE_t by E_p,

$$E_p\sigma_t^2(X_{pt}) \underset{N\to\infty}{=} E_p[(n - X_{pt})/(A - 1)] \ , \quad [12]$$

which is computable from the test scores of a single replication.

The formula for test-retest reliability computable from the test scores of a single replication of a multiple-choice test is thus

$$\rho_p(X_{pt}, X_{pt'}) \underset{N\to\infty}{=} \{\sigma_p^2(X_{pt}) - E_p[(n - X_{pt})/(A - 1)]\}/\sigma_p^2(X_{pt}) \ , \quad [13]$$

where X_{pt} is the corrected-for-guessing score of Equation [9]. If $\overline{X}$ is the mean corrected-for-guessing score and S^2 the corresponding variance on any replication, this formula is rewritable in the simpler form

$$\rho_{XX'} \underset{N\to\infty}{=} 1 - [(n - \overline{X})/S^2(A - 1)] \ , \quad [14]$$

where N is the number of persons, n the number of items, and A the number of options (alternatives) per item.

If the number of options (A) varies from item to item, then, the variance of a sum of independent terms being equal to the sum of the variances,

$$\rho_{XX'} \underset{N\to\infty}{=} 1 - (1/S^2) \sum_A [(n_A - \overline{X}_A)/(A - 1)] \ , \quad [15]$$

where $\overline{X}_A$ is the mean (over people) corrected-for-guessing part score for the n_A items having A alternatives. Although Weitzman (1970) described how

to make A-choice items have a guessing rate equal to $1/A$ (ideal items), trying to achieve this objective for a single value of A may be difficult for all test items. Equation [15] is thus useful in the more easily realizable case of a test in which the value of A and, hence, the guessing rate can vary from item to item. In this case also, with $X_p = \Sigma_A X_{pA}$, where X_{pA} is the part score of person p on all A-choice items, the expected value of X_p over replications is equal to T_p.

Monte Carlo Examples

This section uses Monte Carlo data to illustrate the results of the preceding section. The data consist of computer-simulated responses on two independent occasions by people in 12 different populations to four-choice items in two 50-item tests. For each test and testing occasion, a person's score was computed as the sum of a true and an error component. The 12 populations were constructed to be large or small and to have true components that, when divided by 50, were beta-distributed with low, medium, or high means and small or large standard deviations, according to a factorial design. Computation of the error components differed for the two tests. For each test, the error component of a score with a true component of T was determined cumulatively, starting with zero, by adding one with probability c and zero with probability $1-c$ for each of the remaining $50-T$ items. One of the tests was ideal. For this test, c was equal to 1/4. The other test was nonideal. For this test, c varied from item to item according to a beta distribution with mean equal to 1/4 and standard deviation equal to 1/8, approximately (actual values, to three decimal places, were 0.248 and 0.122, respectively). The beta distribution was chosen for convenience; a mean of 1/4 was chosen because, according to a common view (Davis, 1952), for most multiple-choice tests the average probability of guessing the correct answer to an item is approximately equal to the reciprocal of the number of alternatives per item; and a standard deviation of 1/8 was chosen to allow probabilities of 1/2 to occur occasionally while ensuring that probabilities greater than 1/2 occur only rarely.

For each of the 12 populations, a value of the test-retest reliability, denoted by $\hat{\rho}_{XX'}$, was computed directly from the scores on each test on the two independent occasions. Values of $\rho_{XX'}$ based on Equation [14] and of the Kuder-Richardson Formula 21 (KR-21) were also computed for each population from the scores on each test and testing occasion. These results, together with the mean and standard deviation of the true components and the size of each population, appear in Tables 1 and 2. The data in Table 1 show that $\rho_{XX'}$ is a highly accurate estimator of $\hat{\rho}_{XX'}$ when populations are large; the data in Table 2, by comparison, show that $\rho_{XX'}$ is a somewhat less accurate estimator of $\hat{\rho}_{XX'}$, on the low side, when populations are small. The data in both tables

TABLE 1

CORRELATIONS FOR MONTE CARLO TESTS: LARGE POPULATIONS

μ_T	σ_T	$\hat{\rho}_{XX'}$	Test 1		Test 2		N
			$\rho_{XX'}$	KR-21	$\rho_{XX'}$	KR-21	
			Ideal Tests				
12.3	7.2	.81	.81	.68	.81	.69	976
12.7	9.6	.88	.88	.81	.88	.80	937
25.0	7.3	.87	.87	.68	.87	.69	975
25.0	9.8	.92	.92	.82	.92	.82	970
37.7	7.2	.94	.93	.79	.93	.78	976
37.3	9.6	.96	.96	.88	.96	.87	937
			Nonideal Tests				
12.3	7.2	.82	.80	.67	.81	.68	976
12.7	9.6	.89	.88	.80	.88	.81	937
25.0	7.3	.89	.86	.67	.87	.68	975
25.0	9.8	.93	.92	.82	.92	.82	970
37.7	7.2	.93	.93	.77	.93	.77	976
37.3	9.6	.96	.96	.88	.96	.88	937

show that $\rho_{XX'}$ is a much better estimator of $\hat{\rho}_{XX'}$ than KR-21. This result is noteworthy because, except for the number of options per item, both these statistics are computed from the same data (the total test scores of persons on single testing occasions). Finally, comparison of the upper and lower halves

TABLE 2

CORRELATIONS FOR MONTE CARLO TESTS: SMALL POPULATIONS

μ_T	σ_T	$\hat{\rho}_{XX'}$	Test 1		Test 2		N
			$\rho_{XX'}$	KR-21	$\rho_{XX'}$	KR-21	
			Ideal Tests				
11.5	6.2	.79	.75	.59	.75	.60	84
10.4	7.6	.87	.82	.73	.79	.68	75
25.0	6.7	.87	.83	.61	.84	.61	80
25.0	8.8	.91	.91	.78	.90	.76	79
39.7	6.3	.93	.91	.72	.92	.75	84
39.6	7.6	.96	.95	.84	.95	.84	75
			Nonideal Tests				
11.5	6.2	.80	.78	.64	.71	.53	84
10.4	7.6	.82	.81	.71	.78	.66	75
25.0	6.7	.92	.83	.60	.86	.66	80
25.0	8.8	.91	.90	.76	.89	.75	79
39.7	6.3	.94	.91	.72	.92	.76	84
39.6	7.6	.96	.94	.82	.94	.82	75

of the two tables suggests that in practice $\rho_{XX'}$, as an estimator of $\hat{\rho}_{XX'}$, is likely to be only slightly less accurate for a nonideal than for an ideal test.

Discussion

The single-trial estimators of test-retest reliability presented here (Equations [14] and [15]) require the same assumption as the standard correction for guessing: The guessing rate for an A-choice item is equal to $1/A$. A technology exists for the construction of items that meet this assumption (Weitzman, 1970). The two estimators are asymptotically exact with respect to the number of people tested, unlike other single-trial estimators, which tend to be lower bounds (Guttman, 1945). Users of the standard correction for guessing should also use one of these estimators, as appropriate, to indicate test-retest reliability if the number of people tested is large. National testing programs, in particular, ought to incorporate use of these estimators in conjunction with the standard correction for guessing.

REFERENCES

DAVIS, F. B. Item analysis in relation to educational and psychological testing. *Psychological Bulletin*, 1952, 44, 97-121.

GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.

GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 255-282.

KUDER, C. F., & RICHARDSON, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, 151-160.

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

WEITZMAN, R. A. How the number and relative guessworthiness of alternatives affects the reliability of a multiple-choice test. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, 3, 187-188.

WEITZMAN, R. A. Ideal multiple choice items. *Journal of the American Statistical Association*, 1970, 65, 71-89.

Accepted January 24, 1984.